POPDIST, Version. 1.2.4: User's Guide

Bernt Guldbrandtsen[*], Jürgen Tomiuk[†], and Volker Loeschcke[‡]

October 9, 2012

[*]Dept. of Genetics and Biotechnology, University of Aarhus, Research Center Foulum, PO. Box 50, DK–8830 Tjele, Denmark, Phone: (+45)89991227, Fax: (+45)89991300,e-mail: bg@genetics.agrsci.dk

[†]Dept. of Anthropology and Human Genetics, Wilhelmsstrasse 27, D–4000 Tübingen, Germany, Phone: (+49)(0)7071–297–6883, Fax: (+49)(0)7071–297–5233, e–mail: juergen.tomiuk@uni-tuebingen.de

[‡]Dept. of Genetics and Ecology, Aarhus University, Ny Munkegade, Bldg. 540, DK–8000 Aarhus C, Denmark, Phone: (+45)8942-3268, Fax: (+45)86127191, e-mail: volker.loeschcke@biology.aau.dk

Contents

1 Introduction **2**

2 Input File **2**
2.1 The Header 2
2.2 Data For One Population 2
2.2.1 Missing Values 3
2.3 Polyploid Populations 3

3 Running the Program **5**
3.1 Specifying Input Files 5
3.2 Choosing Measures and Other Options 5
3.3 Using the Menus 5
3.3.1 The File Selection Screen 8
3.3.2 The Measure Selection Screen 9
3.3.3 Measure Selection Commands 9
3.3.4 Measure Variant Selection Commands 10
3.4 Output 10
3.4.1 Output Values 11

4 An Example **11**

5 Obtaining the Program **12**

6 Bugs and Comments **12**

7 Acknowledgements **12**

A Summary of Options **13**

1 Introduction

`POPDIST` is a program for the calculation of various population genetic distance and identity measures. It uses an input file format that is similar to the input format used by the `GENEPOP` software (Raymond & Rousset, 1995), and files created for `GENEPOP` for diploid populations are read unmodified by `POPDIST`.

2 Input File

The genetic data are entered into the program by reading one or more input files. The input file format is an extended version of the input file format for the `GENEPOP` package. The file format is a simple ASCII format. Each file is composed of one header and the genetic data for one or more populations.

The program is executed either through command line options or through a simple menu system. It is the first program to implement the genetic distance and identity measures of Tomiuk & Loeschcke (1991, 1995) for codominant alleles. The measure is very robust against non–equilibrium conditions and can also be applied to microsatellite data. In contrast to alternative measures this measure is able to estimate genetic distances involving polyploid and parthenogenetic populations (Tomiuk & Loeschcke, 1996). Also the measure of Tomiuk and Loeschcke has recently been shown to have very desirable statistical properties (Tomiuk *et al.*, 1998). The program has run quite fast regardless of which of the available options are included, and it is reasonably memory efficient.

2.1 The Header

The header is composed of one line (of comments) followed by lines with the designations of the loci, one name on each line beginning in column 1 of the lines. The first line is discarded by the program, while the locus designations are used in the program output. The number of loci and their names must be consistent across input files if several input files are to be used in the same run. Genetic data are entered population by population. Genetic data for one or more populations are given in one file, but data for one population cannot be contained in more than one file.

2.2 Data For One Population

Data for one population are presented by one line containing the keyword `pop` at the beginning of a line. After that follow lines containing data for one individual each. Each individual line begins with a designation identifying the population that this individual belongs to. This designation for the first individual is used by the program as the label for this population. Designations for subsequent individuals are ignored. The designation field is separated from the genotypes by a comma. To the right of the comma are a number of fields equal to the number of loci. The fields are separated by spaces or tab–characters. Each field

consists of an even number of digits. The number of digits is two times the ploidy level in the population, i.e., in a diploid population there will be 4 digits for each locus, in a triploid 6 digits and so forth. Each pair of digits shows what allele a particular gene belongs to.

An example of the data for a tiny population of diploids that has been typed for two loci with three alleles in each population would be:

```
Pop
Examplpop1      , 0102 0103
Examplpop1      , 0101 0203
```

2.2.1 Missing Values

Missing values are generally specified as `0000` (for a diploid). If one or both populations in a comparison have no data available at a locus a warning will be written. If this is true for all loci for a comparison of two populations this will be indicated by that a string of stars appears in the output instead of a number.

2.3 Polyploid Populations

Polyploid populations are encoded exactly like diploid populations with the exception that the number of digits for each locus for each individual is changed from 4 to 2 times the ploidy level, e.g. 6 digits for a triploid, 8 digits for a tetraploid *etc.*

A simple triploid population might look like:

```
Pop
Examplpop2      , 010101 010103
Examplpop2      , 010202 020203
Examplpop2      , 010101 020203
```

Since in general the number of copies of each allele in a polyploid cannot be observed directly, genotypes must be padded with extra copies of alleles present in the genotype or with zeroes. I.e. the following all for individuals' genotypes each represent the same genotype:

```
Pop
Examplpop2      , 010102
Examplpop2      , 010202
Examplpop2      , 010002
Examplpop2      , 020001
```

Likewise the followin all represent the same genotype:

```
Pop
Examplpop2      , 010000
Examplpop2      , 000001
Examplpop2      , 010101
Examplpop2      , 010001
```

Calculating measures currently only works for the polyploid version of the Tomiuk & Loeschcke measure and the Band Sharing Measure. Hedrick's measure can also be used for comparing polyploid populations of the same level of ploidy. However, this measure is currently not supported directly by the program. It can be calculated assigning each genotype a 4 digit code unique within each locus, and then running the program choosing Hedrick's measure as though the population were diploid.

In this format the triploid population shown above might be coded as:

```
Pop
Examplpop3      , 0101 0101
Examplpop3      , 0102 0102
Examplpop3      , 0101 0102
```

3 Running the Program

The program can be launched from a command line (under MS DOS or UNIX). On the Macintosh it is launched by double clicking the program icon. If you are using a Macintosh skip forward to section 3.3 for a description of the menus.

3.1 Specifying Input Files

Files containing the data for the populations to be examined can be specified on the command line just by specifying their names. Alternatively, they can be specified in a file, one name of an input file on each line. POPDIST is then launched with the `-f` option followed by the name of the file containing the filenames. A third possibility is to use the same set of input files as was used with the last run of POPDIST in the current directory. This is done by specifying the `-O` option. These three possibilities can be combined. Any duplicates will automatically be ignored. When POPDIST terminates, it writes the names of the files used to the file "`oldfiles.tlh`" in the current directory. It is the content of this file that is read when using the `-O` option is used in a later run.

3.2 Choosing Measures and Other Options

The options for choosing among the measures are shown in table 1.

Most measures come in several modifications; often they allow both the estimation of genetic identity between populations and of genetic distance between populations. Some measures have different variants of a distance measure depending on whether reconstruction of topologies or distances (*sensu* Takezaki & Nei (1996)) is desired. Additionally, there is a number of options modifying the actions of the program. These options are shown in table 2.

The capabilities of the measures that have been implemented in POPDIST are summarized in table 3.

3.3 Using the Menus

All the features of the program are also available through a simple keyboard driven menu system. On a Macintosh the program automatically launches with the menus. Under MS DOS and Unix the menus can be launched in either of two ways: Either by just typing `popdist` at the prompt, or by using command line options and including the `-m` option.

Option	Measure
`-n`	Use the measure of Nei (1972)
`-N`	Use the measure of Nei (1978)
`-h`	Use the measure of Hillis (1984)
`-e`	Use the weighted measure of Reynolds *et al.* (1983)
`-u`	Use the unweighted measure of Reynolds *et al.* (1983)
`-g`	Use the measure of Goldstein *et al.* (1995)
`-c`	Use the measure of Hedrick (1971)
`-a`	Use the measure of Cavalli-Sforza & Edwards (1967)
`-t`	Use the measure of Tomiuk & Loeschcke (1991, 1995) (default).

Table 1: Summary of options for choosing genetic measures implemented in POPDIST.

Option	Version of Measure
`-i`	Use a identity estimating measure (if available)
`-d`	Use a distance reconstructing measure (if available)
`-p`	Use a topology reconstructing measure (if available)
	Other Functions
`-m`	Start up with menu
`-j`	Calculate jackknife of standard error of the estimate over loci
`-O`	Reuse the files in the previous run (as stored in the file "`oldfiles.tlh`" in the current directory)
`-f filename`	Use files specified in the file given by filename
`-o filename`	Put out put into the file given by filename instead of to the screen
`-s screenwidth`	Set width of screen for output to screenwidth characters for output

Table 2: Summary of auxiliary functions and variants of measures implemented in POPDIST.

Measure	Property			
	Topology	Distance	Identity	Polyploid
Goldstein *et al.* (1995)		•		
Hedrick (1971)			•	•[1]
Hillis (1984)		•	•	
Nei (1972)	•[2]	•	•	
Nei (1978)		•	•	
Reynolds *et al.* (1983) (weighted)		•	•	
Reynolds *et al.* (1983) (unweighted)		•	•	
Rogers (1972)		•		
Cavalli-Sforza & Edwards (1967)		•		
Tomiuk & Loeschcke (1991)	•[3]	•	•	•
Band Sharing measure		•	•	•

Table 3: Summary of the capabilities of the measures implemented in `POPDIST` version 1.2.4. Topology and distance both indicate measures of genetic distance. If both Topology and distance are indicated they are the different recommended forms of the measure for reconstructing tree topology and genetic distances. Identity indicates whether the method implements a measure of genetic identity. Polyploid indicates that the measure is applicable to comparisons involving polyploid populations. 1. The polyploid version of Hedrick's measure is only indirectly supported (see text). 2. Described by Nei *et al.* (1983) and Takezaki & Nei (1996). 3. The distance reconstructing measure without the square root correction (Tomiuk *et al.*, 1998).

The menu has two screens, one for entering (and removing) data files from the calculation, and one for choosing the measure to be calculated. Everything operates by typing one–letter commands followed by hitting the "return" key (which be labeled "enter" depending on your keyboard). Depending on the command you gave, the program may then prompt you for additional data. Type in the data and finish by hitting "return". All commands are case insensitive, i.e. uppercase and lowercase letters do the same thing.

3.3.1 The File Selection Screen

The File Selection screen consists of three parts:

- First it lists the names of input files entered up to now. I you started the program on a Macintosh or by just typing `popdist` followed by "return", it should just say "`No files`" here at this point.
- The second part shows where output will be written. The default is to write output to the screen.
- The third part shows the available commands.

The available commands are:

The `a` command The simplest way to add files is to use the `a` command. You will be prompted for the name of an input file in the current directory. Files will be added to the list of data files selected.

The `r` command To remove a file from the list of input files use the `r` command. You will be prompted for the number of the file to remove. The numbers of the files currently selected are shown in the list of input files entered up to now.

The `f` command To make life easier if you want to calculate several measures for the same set of data files put the names of the data files into a file, one name per line. If you then use the `f` command you will be prompted for the name of such a file with data files.

The `p` command If you just want to use the same data files as in the previous run, use the `p` command. It will cause POPDIST to read the file "`oldfiles.tlh`" in the current directory. This file was written at the end of the last run in this directory. If the program hasn't been run in this directory before (i.e. the file "`oldfiles.tlh`" does not exist), the command does nothing.

The `o` command Giving the `o` command will make the program prompt you for the name of a file where the genetics distance or similarity measures are to be written.

The `q` command Using this command causes the program to exit immediately.

The `?` command This will show a brief summary of the commands.

The `c` command When you are done with selecting data files and optionally an output file, use the `c` command to continue to the next screen.

3.3.2 The Measure Selection Screen

The Measure Selection screen consists of two parts:

- First the current choices are shown. The choices consist of three things: First the measure currently chosen, second whether Jackknife of the standard error of the estimates will be calculated, and third what variant of the measure is calculated. Notice that many combinations of choices measure and variant are valid. See table 3 for valid combinations.
- Second there is a listing of the available commands.

3.3.3 Measure Selection Commands

The `t` command chooses the measure of Tomiuk & Loeschcke (1991). This is the default.

The `n` command chooses the measure of Nei (1972).

The `r` command chooses the measure of Rogers (1972).

The `e` command chooses the weighted version of the measure of Reynolds *et al.* (1983).

The `h` command chooses the measure of Hillis (1984).

The `g` command chooses the measure of Goldstein *et al.* (1995).

The `k` command chooses the measure of Hedrick (1971).

The `7` command chooses the measure of Nei (1978).

The `w` command chooses the unweighted version of the measure of Reynolds *et al.* (1983).

The `a` command chooses the chord version of the measure of Cavalli-Sforza & Edwards (1967). Notice, that I divide the measure by the number of loci, in order to handle cases where comparisons between pairs of populations differ in the number of loci for which data are available for comparison, i.e. when one (or both) populations in a comparions have no data for one or more loci.

The `b` commnand chooses the band shareing measure of Tomium *et al.* (*unpubl.*)

3.3.4 Measure Variant Selection Commands

The `p` command chooses the topology reconstructing version of a genetic distance measure (if available).

The `s` command chooses the distance reconstructing version of a genetic distance measure (if available).

The `i` command chooses to do estimation of genetic identity. This is the default.

The `j` command chooses to do jackknife estimation of the standard error.

The `o` command chooses not to do jackknife estimation of the standard error. This is the default.

The `?` command This will show a brief summary of the commands.

The `q` command Using this command causes the program to exit immediately.

The `c` command causes the program to calculate genetic measures, as currently selected.

3.4 Output

The program returns a matrix of genetic identities or, optionally, genetic distances between the set of populations described in the data file(s). The program can calculate jackknife of standard error. If the jackknife option is used the mean of the individual jackknife estimates is calculated instead of the normal joint estimate across all loci. Note, however, that the last two estimates may be identical, depending on the measure chosen.

The width of the matrix defaults to 80 characters. This can be changed with the `-s` option (minimum 30).

Normally the output is sent to the screen. However this can be modified using the `-o` option, which can be used to specify an output file.

3.4.1 Output Values

If the `-j` option is not used, i.e. jackknife estimates are not calculated, all output values are the values given by the measure in their respective standard way. If, on the other hand, the `-j` option is use, then the uncorrected mean of the individual jackknife estimates are given, in each case this number is then followed by the jackknife standard error of the mean.

In some cases some comparisons may have no data. In these cases the output values are replaced by `*****`. Also, comparisons may result in impossible values, e.g. log(0). These are also represented by `*****`.

4 An Example

One data file called `exmpl1.dat` will be used. It contains the data shown below:

```
This is example 1
Enzyme1
Enzyme2
Pop
Examplpop1      , 0102 0103
Examplpop1      , 0101 0204
Pop
Examplpop2      , 0101 0103
Examplpop2      , 0102 0203
Examplpop2      , 0104 0203
```

To calculate the Tomiuk & Loeschcke genetic distance use the command `popdist exmpl1.dat`. The output should then be:

```
popdist version  1.0, built Fri Apr 17 11:04:53 NFT 1998
Tomiuk & Loeschcke's Identity estimating measure
Population | Examplpop1 Examplpop2
-------------------------------------------------------------------------------
Examplpop1 | 1.000      0.896
Examplpop2 | 0.896      1.000
```

apart from the version information and the date. To use instead the distance reconstructing measure of Nei (1972), including calculation of the jackknife of standard errors use the command: `popdist -n -d -j exmpl1.dat` to obtain:

```
Nei(72)'s distance reconstructing measure
Population | Examplpop1     Examplpop2
-------------------------------------------------------------------------------
Examplpop1 | 0.000( 0.000)  0.126( 0.067)
Examplpop2 | 0.126( 0.067)  0.000( 0.000)
```

Next, an example involving one diploid population and one triploid population with the following data in the file exmpl2.dat:

```
This is example 2
Enzyme1
Enzyme2
Pop
Examplpop1      , 0102 0103
Examplpop1      , 0101 0204
Pop
Examplpop2      , 010101 010103
Examplpop2      , 010202 020203
Examplpop2      , 010104 020203
```

When genetic identities with jackknife estimates of the standard deviations using the Tomiuk & Loeschcke measure (with the command `popdist -j exmpl2.dat`), the result should be:

```
Tomiuk & Loeschcke's Identity estimating measure
                 Examplpop1      Examplpop2
-------------------------------------------------
Examplpop1      | 1.000( 0.000)   0.901( 0.025)
Examplpop2      | 0.901( 0.025)   1.000( 0.000)
```

5 Obtaining the Program

The program can be obtained from http://genetics.agrsci.dk/ bernt/popgen. Users who do not have access to the World Wide Web may send a MS DOS–formatted 1.4 MB disk to the first author. The program will be available in precompiled versions for the Macintosh, IBM PC–compatible computers, and various brands of UNIX. Availability of UNIX precompiled versions depends on what machines currently are available to us. Up–to–date information about new versions of the program will be available at the web site.

6 Bugs and Comments

If you discover any problem due to the program, however minor, we would like to hear about it. Please send e-mail to mailto:bernt.guldbrandtsen@agrsci.dk.

Also, if you succeed at compiling the program on platforms for which it is currently not available on the web page mentioned in section 5, we would appreciate a copy, so that it can be made available on the net.

7 Acknowledgements

We thank the European Science Foundation for supporting the visit of JT to Aarhus (grant no. ESF-POBI/95Y) and the Danish Natural Science Research council for supporting parts of the project (grant no. 9701412 to VL).

A Summary of Options

`-7`	Use the measure of Nei (1978)
`-b`	Use the measure of Tomiuk *et al.* (*unpubl.*)
`-c`	Use the measure of Hedrick (1971)
`-d`	Use a distance reconstructing measure (if available). If a distance reconstructing version of the measure is not available the program chooses another measure to calculate
`-e`	Use the weighted measure of Reynolds *et al.* (1983)
`-f filename`	Use files specified in the file given by filename
`-g`	Use the measure of Goldstein *et al.* (1995)
`-h`	Use the measure of Hillis (1984)
`-i`	Use a identity estimating measure (if available) If a identity reconstructing version of the measure is not available the program chooses another measure to calculate
`-j`	Calculate jackknife of standard error of the estimate over loci
`-m`	Start up with menu
`-n`	Use the measure of Nei (1972)
`-O`	Reuse the files in the previous run (as stored in the file "`oldfiles.tlh`" in the current directory)
`-o filename`	Put out put into the file given by filename instead of to the screen
`-p`	Use a topology reconstructing measure (if available). If a topology reconstructing version of the measure is not available the program chooses another measure to calculate.
`-s screenwidth`	Set width of screen for output to screenwidth characters for output
`-t`	Use the measure of Tomiuk & Loeschcke (1991, 1995)
`-u`	Use the unweighted measure of Reynolds *et al.* (1983)

References

Cavalli-Sforza, L.L., & Edwards, A.W.F. 1967. Phylogenetic Analysis: Models and Estimation Procedures. *Evolution*, **21**(September), 550–570.

Goldstein, D.B., Ruis Linares, A., Cavalli-Sforza, L.L., & Feldman, M.W. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463–471.

Hedrick, P.W. 1971. A new approach to measuring genetic similarity. *Evolution*, **25**, 276–280.

Hillis, D.M. 1984. Misuse and modification of Nei's genetic distance. *Syst. Zool.*, **33**, 238–240.

Nei, M. 1972. Genetic distance between populations. *Amer. Natur.*, **106**, 283–292.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.

Nei, M., Tajima, F., & Tateno, Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.*, **91**, 153–170.

Raymond, M., & Rousset, F. 1995. GENEPOP (version 1.2): a population genetics software for exact tests and ecumenicism. *J. Heredity*, **86**, 248–249.

Reynolds, J., Weir, B.S., & Cockerham, C.C. 1983. Estimation of the coancestry coefficient: Basis for a short–term genetic distance. *Genetics*, **105**, 767–779.

Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. *Pages 145–153 of: Studies in Genetics VII.* Univ. Texas Publ. 7213.

Takezaki, N., & Nei, M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, **144**, 389–399.

Tomiuk, J., & Loeschcke, V. 1991. A new measure of genetic identity between populations of sexual and asexual species. *Evolution*, **45**, 1685–1694.

Tomiuk, J., & Loeschcke, V. 1995. Genetic identity combining mutation and drift. *J. Heredity*, **74**, 607–615.

Tomiuk, J., & Loeschcke, V. 1996. A maximum–likelihood estimator of the genetic identity between polyploid species. *J. Theor. Biol.*, **179**, 51–54.

Tomiuk, J., Guldbrandtsen, B., & Loeschcke, V. 1998. Population differentiation through mutation and drift – A comparison of genetic identity measures. *Genetica*, **102/103**, 545–558.